一、数据预处理

1. 处理缺失值:

对数据标签为 location 和 country 的数据使用 unknown 来填补缺失值，对数值型的数据 total_influence 使用中位数填充。

2. 删除重复数据:

3. 数据类型的统一:

将 total_influence 统一转化成浮点数的格式，event_time 应为时间类型。

4. 数据异常处理:

对 total_influence 中数值过于大的数据（中位数超过 0.99）进行处理。

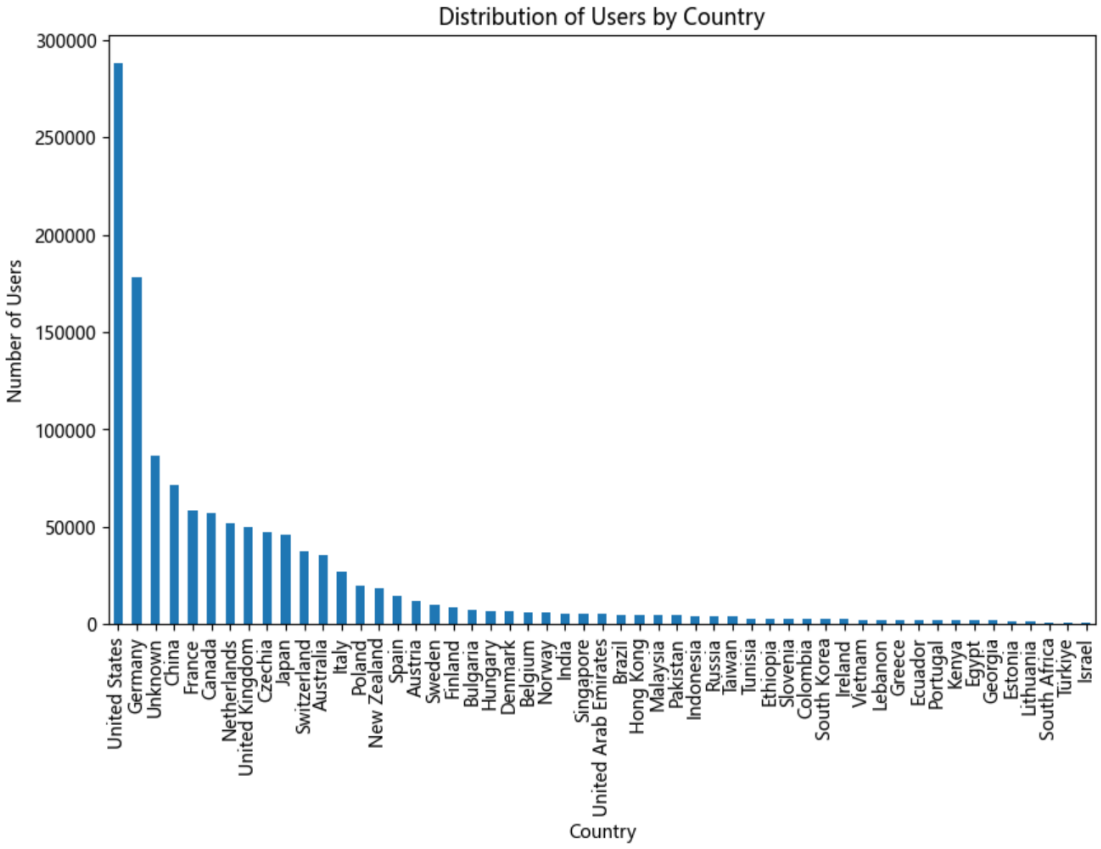
5. 标准化:

对 location 和 country 字段去除前后空格并统一大小写。

二、数据探索

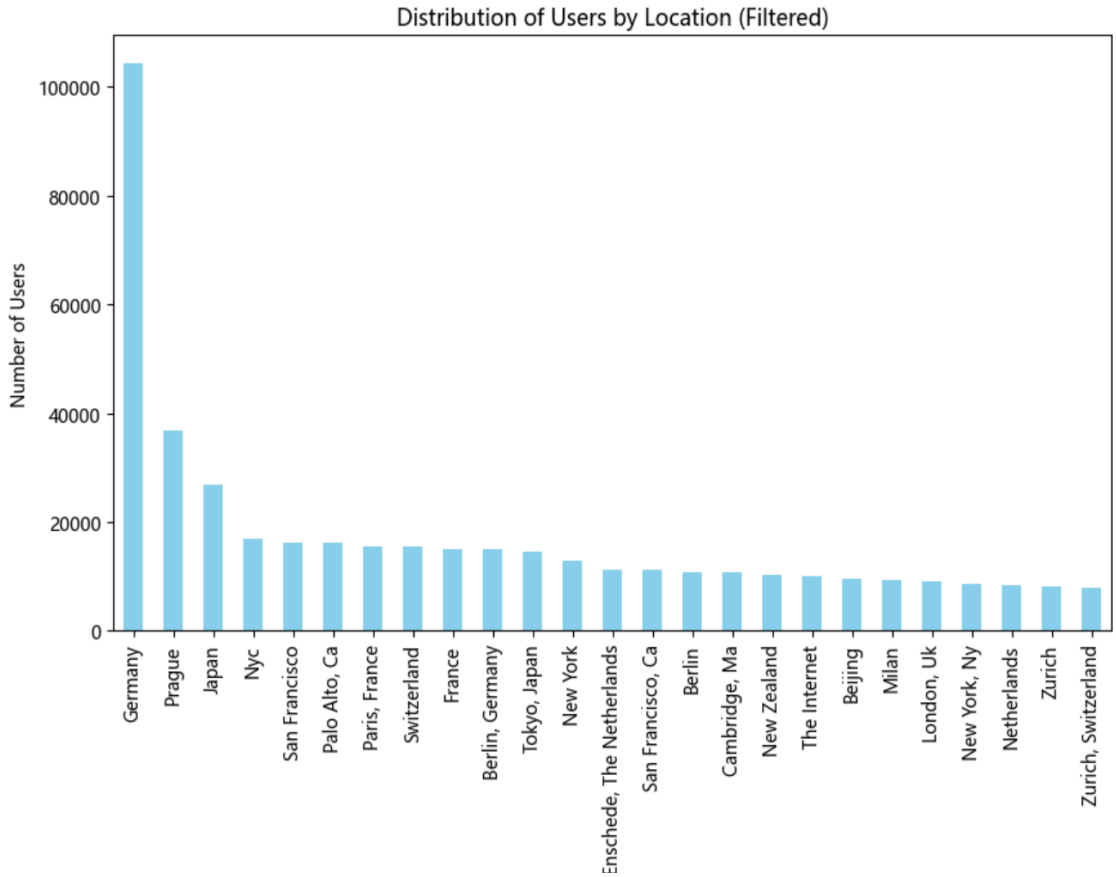
1. 人口统计分析

主要国家分布图:



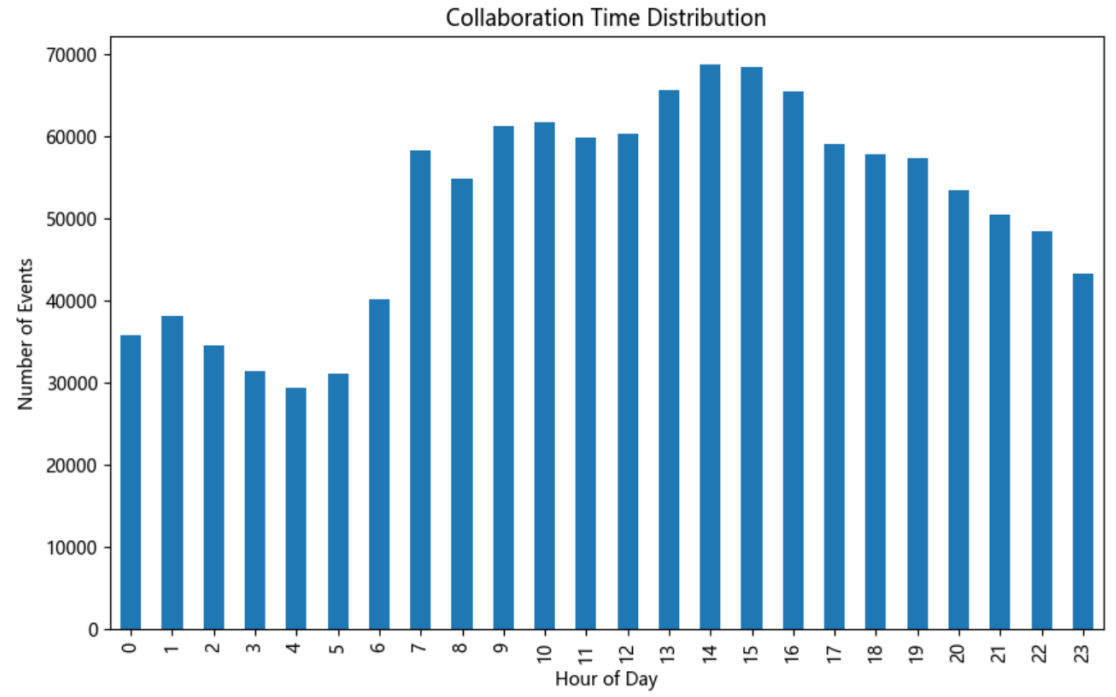
由图可以看出，开发者主要聚集地在 US，其次是 Germany，中国其次，其中还可以看出有很多缺失值 known 的数据。

城市分布图:

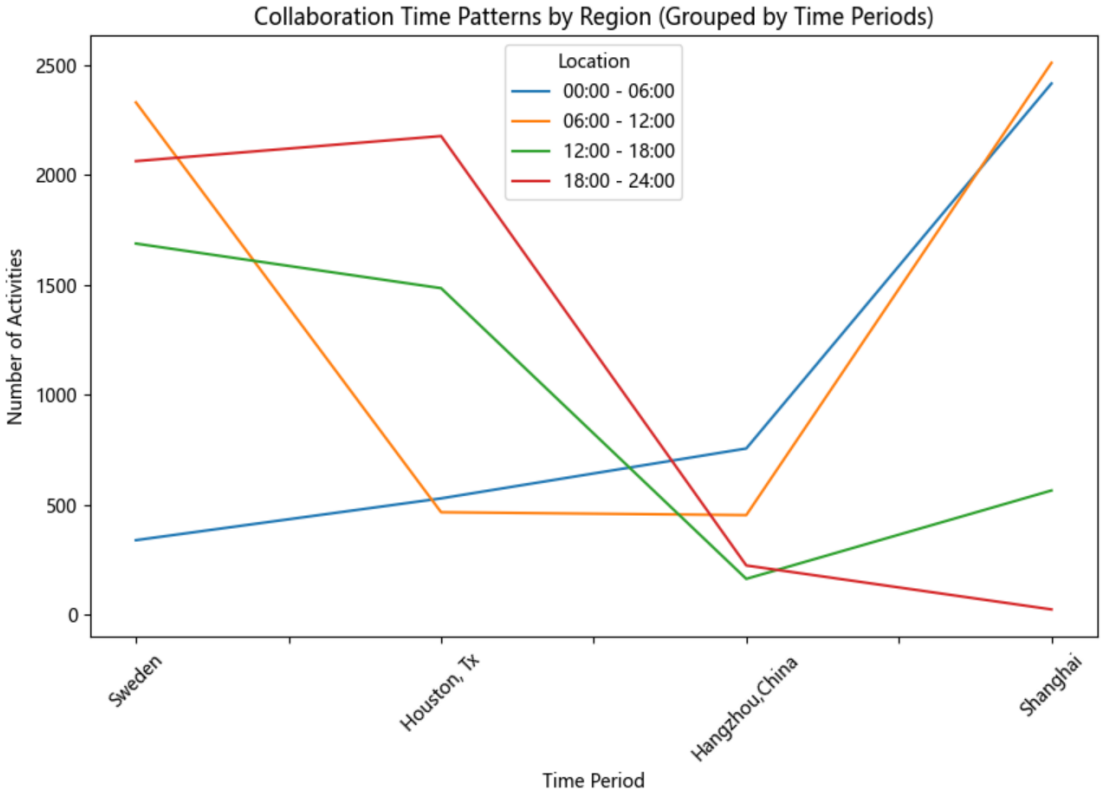


从图中可以得知用户主要城市分布与主要国家分布基本呈现大致的状态，Germany 用户还是居于最高，并且远远超过第二名 Prague，其余城市如 San Francisco, Nyc, Japan 等国家的用户数量大致相似，没有很大的落差。

时间分布：

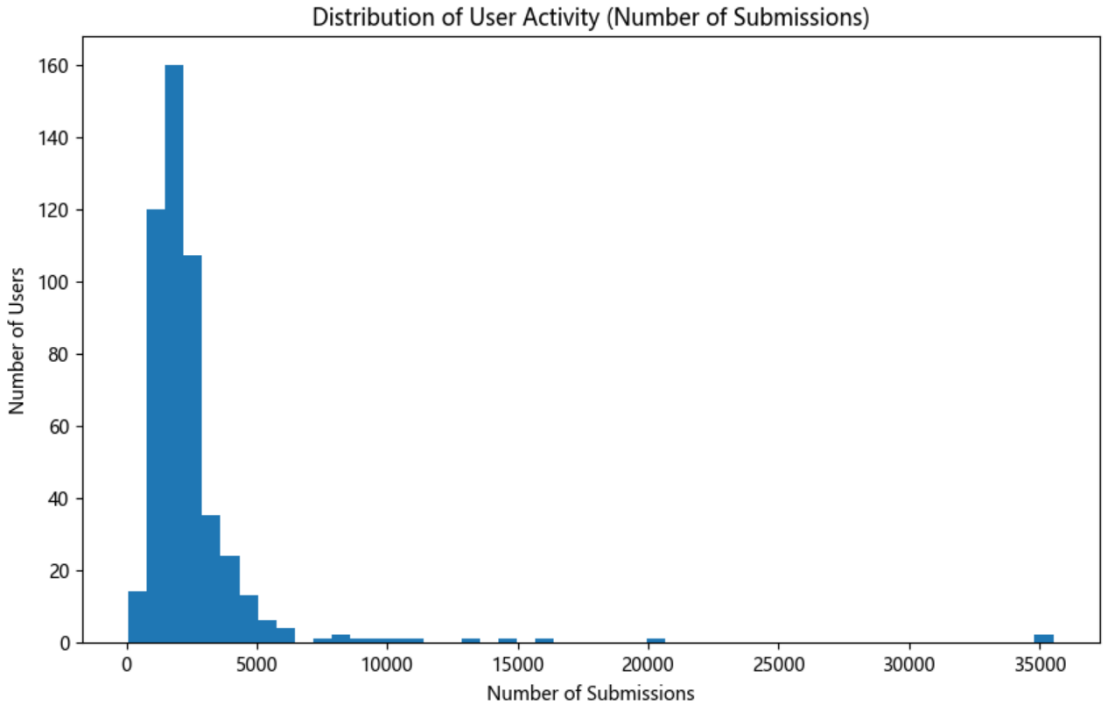


从图中可以看出，用户主要活跃的时间是 7 点到 15 点，但是每个时间段都有，并且差距相差并不是特别大，这是因为用户的协作时间都被转化成了我们这个时区，在我们的凌晨可能是别的时区的白天，所以凌晨的协作量并没有很低。



上面这张图片是选取了两个西方国家和两个国内城市的工作时区分布，从中对比可以看出西方国家的工作区间和国内的相差较大，这是因为地区与地区之间的时区并不相同，统计的时间是北京时间，此外，从用户的主要协作时间区间可以推断这个用户大概是哪个时区，比如（在北京时间下）6 点到 12 点工作的时区大概率是国内，而 18 点到 24 点工作的时区大概率是国外的某些城市。从图中可以验证这一点。

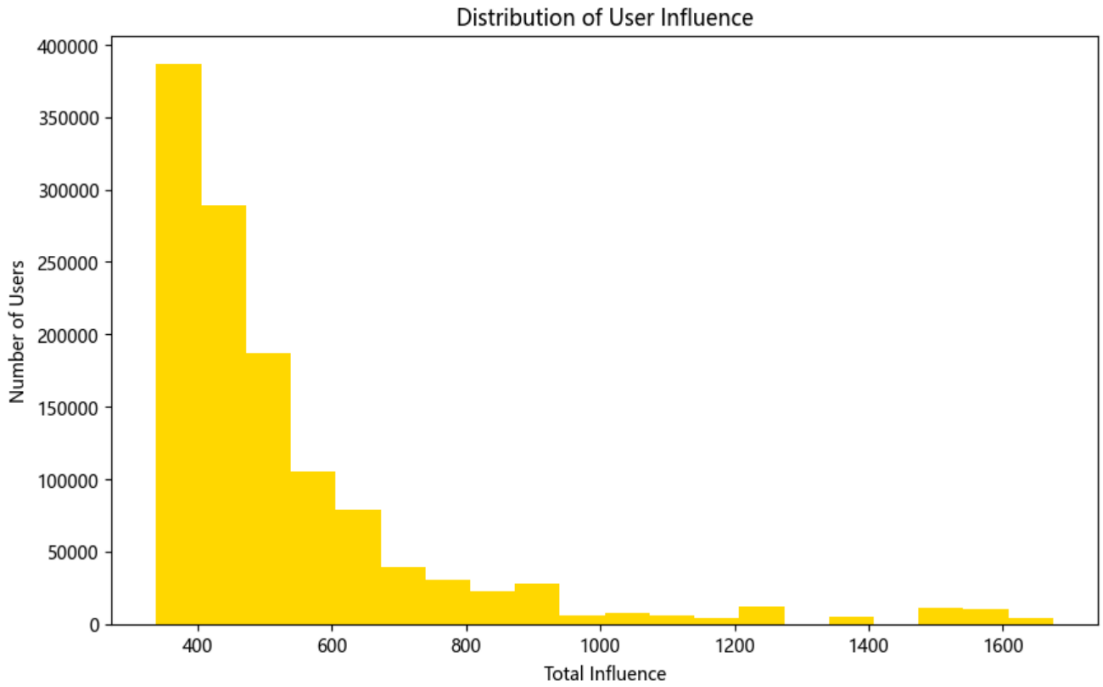
2. 协作行为分析
   提交频率图:



从图中可以获知大多数用户提交的频率在 0 到 5000 次之间，很少数的用户可以达到 10000 次左右，通过使用中位数来区分高活跃和低活跃度用户，百分之 10 为高活跃度，后百分之 10 为低活跃度用户可以清晰被发现：

高活跃用户：user_id		低活跃用户：user_id	
11146458	35532	3705199	1052
158862	35237	446555	1043
1580956	20214	914682	1025
40306929	16149	2536374	1025
43724913	14492	16336606	1017
9824526	12892	432549	1008
195327	11185	26427004	1005
28706372	10448	38668450	1001
50149701	9331	1794099	992
20182680	9095	1221575	972
138339	8300	34168	967
8188402	8244	1096616	960
46537034	7499	24123821	950
19890804	6428		

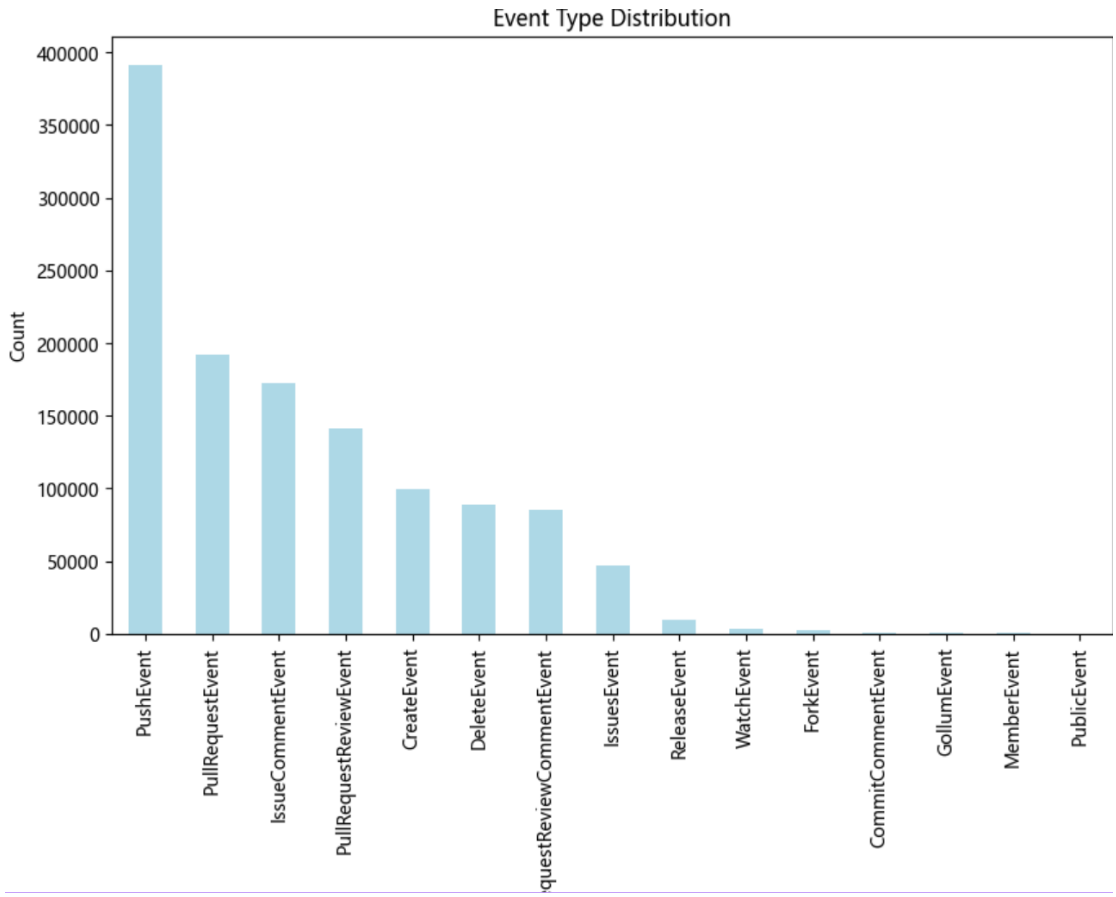
3. 分析影响力和活跃度的关系



从图中可以得出，影响力的趋势和活跃度的趋势大致存在相似之处，于是考虑影响力和活跃度之间是否存在某些关系或者会不会相互影响。

首先可以分析一下所有用户的平均影响力，接着将它和高活跃度用户的平均影响力对比可以发现，一般情况下，活跃度越高的用户其影响力会偏高

4. 统不同事件类型的数量:



从图中可以看出，用户做的最多的操作是 Push，其数量远超过 Pullrequest 和 comment，从此处可以说明大多数用户倾向于对仓库的直接修改，comment 数量远远小于 push 说明用户可能在 push 的时候不喜欢进行 comment，可能是个人偏好，也可能是没有在进行多人的协作项目。